

EXPRESS MAIL number: ET014038726US

Date of Deposit: August 7, 2001

I hereby certify that this paper is being deposited with the United States Postal Service "EXPRESS MAIL Post Office to Addressee" service under 37 CFR 1.10 on the date indicated above and is addressed to the Assistant Commissioner for Patents; Washington, DC 20231.

Kelli J. Withrow

Name of person mailing paper or fee



Signature

=====

APPLICATION FOR UNITED STATES LETTERS PATENT

=====

Title: RENDERING DEFORMABLE 3D MODELS RECOVERED FROM VIDEOS

Inventors: Matthew Brand

092006-00001

## Rendering Deformable 3D Models Recovered from Videos

### Field of the Invention

The present invention relates generally to the field of computer vision, and particularly to recovering 3D models including shape, motion, and deformation of non-rigid objects from videos so that videos can be rendered in a novel manner.

### Background of the Invention

The problem of recovering 3D deformable models of non-rigid objects from a video is of intense interest in the field of computer vision. Linear models of variability are particularly desirable. For example, eigenface models have been widely used to model 2D image variability since the 1980's. Eigenface models use a variance-reducing dimensionality-reduction for coding and decoding face images, see U.S. Patent No. 5,164,992 "*Face recognition system*" issued to Turk, et al. on November 17, 1992. Eigenface methods describe variation between images but do not shed any light on the 3D structure of scenes and objects from which the images were generated.

A first class of methods addresses special cases of the recovery problem that are well-constrained by additional information. For example, depth estimates are available from multi-camera stereo cameras or laser range-finders; the objects are rigid; object surfaces are specially decorated with textures or markers to make inter-image correspondences obvious; or structured light is used to reveal contours of the object. These constrained methods require various combinations of high-

quality, high-resolution videos, calibrated cameras, special lighting, and careful posing.

A second class of methods relaxes image constraints but nevertheless depends on having a pre-computed set of possible models or motion, see Blanz et al., “A morphable model for the synthesis of 3D faces,” Proc. SIGGRAPH99, 1999, Bregler et al., “Non-rigid 3D shape from image streams,” Proc. CVPR, 2000. However, they do not address the case of both unconstrained motion and no prior understanding of the shape and motion of the object in the video.

Therefore, there is a need to recover 3D models from non-rigid objects in unconstrained videos so that the models can be used to generate an entirely new video where the objects are posed and deformed in novel ways.

## Summary of the Invention

The present invention provides linear methods for recovering of 3D shape, motion, and deformation of non-rigid 3D objects from a video. The video can be low-quality and low-resolution, such as a home-made video, a low-end consumer video, or a video obtained under poor lighting conditions. The output is a 3D deformable model, and 3D motion and deformation parameters. The parameters then can be modified and applied to the model to render the video in a novel manner.

Factoring the 3D deformable model from 2D correspondences in the input video according to the invention is a quartic optimization problem, and is based on

cascaded single values matrix decompositions (SVD) that generalize prior art rigid object structure-from-motion factorization.

The invention provides a “parsimony” constraint, and uses the constraint to correct  
5 any undesirable properties of the SVDs. In addition, the invention provides improved and generalized methods for factorization of correspondences or intensity variations with respect to uncertainty in the video, and recovers the models directly from image intensities in which inter-image correspondences are found during the factorization.

More particularly, the invention provides a system and method for recovering a non-rigid 3D model of a scene in an input video acquired by a camera. A set of features in the scene are identified in a reference image of the input video. Correspondences between the set of features in the reference image and  
10 corresponding features in each other image of the input video are determined. The correspondences are factored into a motion matrix and a shape matrix from which the non-rigid 3D model is extracted. The 3D model includes a linear basis for deformable shape in the scene, and for each image a 3D rotations matrix, deformation coefficients, and translation vectors. The non-rigid 3D model and the  
15 input video can now be manipulated to generate a novel video for display from the input video.

## Brief Description of the Drawings

Figure 1 is a flow diagram of a method for recovering a non-rigid model of an object in a video;

5

Figure 2 is a flow diagram of a step of the method of Figure 1 for determining image features correspondences;

Figure 3 is a flow diagram of a step of the method of Figure 1 for factoring the feature correspondences into motion and shape matrices; and

Figure 4 is a flow diagram of a step of the method of Figure 1 for extracting a non-rigid 3D from the motion and shape matrices.

Figure 5 is a block diagram of optical flow in successive images of a video;

Figure 6 is graph of recovered motion parameters according to the invention; and

Figure 7 is a graph comparing recovery methods.

20

## Detailed Description of the Preferred Embodiment

### Method Overview

5 Figure 1 shows a method 100 according to my invention. The method 100 recovers a non-rigid 3D shape and motion model 131 of an object or a scene directly from an input video 101 obtained by a camera. For example, the video is of a “talking head” 103 seen face on. Techniques and machinery for acquiring images by video or movie cameras are well known.

First, a reference image 102 is selected. The reference image 102 can be any image in the video that is consistent with the general shape and appearance of the object of interest throughout the video, for example, the middle image. Correspondences 111 are then determined 200 for a set of features 105 of the object in each image of the video.

The features 105 can be selected manually or automatically. For example, a pointing device 106 can be used to “click” on a relatively small number of features in the reference image 102. The number of features depends on the desired amount of detail in the final 3D model. For faces, 80-100 features are often sufficient. For best results, the selected features are in regions wherein image intensities vary substantially. This can also be done automatically using known computerized vision systems and procedures that locate regions where the intensity gradient varies in all directions, for example, corners and dots. For example, for a face, the features can be the corners of the mouth, eyes, tip of the nose, etc. The invention can also find correspondences for image regions that have little or no texture.

Alternatively, the reference image can be partitioned into equal size patches, and a

suitable feature is selected in each patch, or the feature is merely an approximate central point in the patch. This is useful where it is desired to model the entire scene in the video.

- 5 The correspondences include at least the 2D (pixel) location  $(x, y)$  of each feature. The correspondences can also include information about the image intensities gradients around the features. For tracking purposes, the features can generally be characterized by a small number of pixels, e.g.,  $5 \times 5$ , centered at the each feature location. Below, I will describe procedures that recover the 3D model either from estimated feature location ( $P$ ), or temporal image intensity gradients ( $Y$ ), which are measurements used to calculate optical flow.

10  
15  
20  
25  
30  
35  
40  
45  
50  
55  
60  
65  
70  
75  
80  
85  
90  
95  
100  
105  
110  
115  
120  
125  
130  
135  
140  
145  
150  
155  
160  
165  
170  
175  
180  
185  
190  
195  
200  
205  
210  
215  
220  
225  
230  
235  
240  
245  
250  
255  
260  
265  
270  
275  
280  
285  
290  
295  
300  
305  
310  
315  
320  
325  
330  
335  
340  
345  
350  
355  
360  
365  
370  
375  
380  
385  
390  
395  
400  
405  
410  
415  
420  
425  
430  
435  
440  
445  
450  
455  
460  
465  
470  
475  
480  
485  
490  
495  
500  
505  
510  
515  
520  
525  
530  
535  
540  
545  
550  
555  
560  
565  
570  
575  
580  
585  
590  
595  
600  
605  
610  
615  
620  
625  
630  
635  
640  
645  
650  
655  
660  
665  
670  
675  
680  
685  
690  
695  
700  
705  
710  
715  
720  
725  
730  
735  
740  
745  
750  
755  
760  
765  
770  
775  
780  
785  
790  
795  
800  
805  
810  
815  
820  
825  
830  
835  
840  
845  
850  
855  
860  
865  
870  
875  
880  
885  
890  
895  
900  
905  
910  
915  
920  
925  
930  
935  
940  
945  
950  
955  
960  
965  
970  
975  
980  
985  
990  
995

Next, the correspondences 111 are factored 300 using a cascade of singular value matrix decompositions (SVD) and re-orthogonalizations described in greater detail below. The factoring 300 yields a motion matrix ( $M$ ) 121, and a shape matrix ( $S$ ) 122. In contrast with the prior art, the shape matrix 122 is deformable, and therefore can represent a non-rigid object or scene. In contrast with the prior art on estimating deformable models, the shape matrix 122 makes minimal use of deformations to account for the image data. This maximizes the probability that the recovered shape and deformations are physically correct. As described in detail below, the factoring can either use the location of the features, or the intensity measurements used to calculate the optical flow at the feature locations.

The 3D model 131 is then extracted 400 from the matrices 121-122 in such a way that the model conforms to the correspondences of the set of features with a minimum residual error and minimal deformations. This is accomplished by a rank reduction in the motion and shape matrices. As described below in greater detail,

the 3D model including a linear shape matrix  $\mathbf{S}$ , and for each frame a 3D rotations matrix  $\mathbf{R}$ , a deformation coefficients  $\mathbf{c}$  vector, and a translation vector  $\mathbf{T}$ .

The 3D model 131 can now be used to generate an entirely novel video 141 according to user supplied deformation parameters 142. For example, the video 101 of the “talking head” 103 seen front-faced is transformed into a video where the talking head is seen in profile 143.

### Image Formation Model

I use a simple model of image formation to solve the problem of non-rigid, unconstrained model recovery. An measured pose  $\mathbf{P}$  of an object (or scene) in an image is a projection of a sum of weighted shape/deformation basis  $\mathbf{S}$ , rotated  $\mathbf{R}$  in 3D by  $\mathbf{R}$ , and translated by  $\mathbf{T}$  in the image plane, for details, see U.S. Patent Application Sn., 09/791,117 “*Modeling shape, motion, and flexion of non-rigid 3D objects in a sequence of images*,” filed by Brand on February 22, 2001, incorporated herein by reference.

For the input video 101, I can write the projection  $\mathbf{P}_f$  in image (frame)  $f$  as:

$$\mathbf{P}_f = (\mathbf{c}_f^T \otimes \mathbf{R}_f) \mathbf{S} + \mathbf{T}_f, \quad (1)$$

where  $\mathbf{c}^T$  represent the transposed vector of deformation coefficients, and  $\otimes$  signifies the Kronecker product. For multiple images  $\mathbf{F}$ , the model for the pose of the object over time can be written as:

$$\mathbf{P} = \mathbf{M} \mathbf{S} + \mathbf{T}, \quad (2)$$

where  $\mathbf{M}$  is a motion matrix defined by vertically stacking  $(\mathbf{c}_f^T \otimes \mathbf{R}_f)$  for all frames.

Of special interest is the structure of the motion matrix  $\mathbf{M}$ :



$$\mathbf{M} = (\mathbf{C}^T \otimes \mathbf{1}_{D \times 3}) \bullet (\mathbf{1}_{K \times 1} \otimes \mathbf{R}), \quad (3)$$

where  $\mathbf{C}$  captures the temporal evolution of the deformation,  $\bullet$  is the Hadamar product,  $\mathbf{1}$  denotes ones matrices,  $D$  the dimensionality of correspondences, and  $K$  is the number of modes of deformation, including scale changes.

5

It is desired to recover the variables  $\mathbf{C}$ ,  $\mathbf{R}$ ,  $\mathbf{S}$ , and  $\mathbf{T}$  directly from the inter-image correspondences 111 or intensity variations measured in the video 101.

 095036-06934  
T03130-92022640

If all features of the object are measured in all images of the video, the translation  $\mathbf{T}$  can be estimated as the mean of  $\mathbf{P}$ , and then removed from  $\mathbf{P}$  to give a matrix  $\underline{\mathbf{P}}$  whose rows are zero-mean. A SVD of  $\underline{\mathbf{P}}$  yields the motion matrix  $\mathbf{M}$  121, and the basis  $\mathbf{S}$  122. Then, the motion matrix  $\mathbf{M}$ , in turn, can be decomposed into rotations  $\mathbf{R}$  and deformation coefficients  $\mathbf{C}$ . The details of these decompositions follows below.

15

Like many multi-linear phenomena in images, a key to the successful factorization 300 of the image data, i.e., a factorization that is consistent with the forward model in equation (1), is the application of rank and norm constraints to substructures in these matrices.

20

### Rigid Object Factorization

25

In the rank  $K = 1$  case of rigid object motion, a rank-3 thin SVD  $\mathbf{MS} \stackrel{\text{SVD}}{\leftarrow} \underline{\mathbf{P}}$  factors motion and shape information from rigid tracking data, see Tomasi et al., “*Shape and motion from image streams under orthography: A factorization method*,” International Journal of Computer Vision, 9(2):137–154, 1992. The pseudo-motion matrix  $\mathbf{M}$  of left singular vectors associated with the three largest singular values

5

The SVD determines both sides up to an invertible 3D affine transformation  $\mathbf{G}_{3 \times 3}$ , such that  $\mathbf{MS} = \mathbf{MG}^{-1}\mathbf{GS} = \mathbf{P}$ . In order to get proper rotations and shape, one solves for a  $\mathbf{G}^{-1}$  that restores orthogonal structure to the motion matrix  $\mathbf{M}$ . The orthogonality of the row vectors  $\mathbf{m}$  of the  $x$  and  $y$  components of the rotation matrix  $\mathbf{R}$  in the matrix  $\mathbf{M}_f$  in image  $f$  give a constraint that is linear in six unknowns of a symmetric matrix  $\mathbf{H} = \mathbf{G}^{-1}\mathbf{G}^{-T}$ . This can be obtained via standard least-squares methods from a system of linear constraints. Because the matrix  $\mathbf{H}$  is symmetric, the constraints on the matrix can be expressed very concisely, and an approximate  $\mathbf{G}^{-1}$  can be determined from a SVD on the matrix  $\mathbf{H}$ .

20 A proposal to extend the above factorization to non-rigid objects as described by Bregler et al. “*Recovering non-rigid 3D shape from image streams*,” Proc. CVPR, 2000 (BHB factorization). For rank  $K$  deformation bases, one performs a SVD of  $\mathbf{P}$  and retains the top ranked three singular vectors. The shape basis  $\mathbf{S}$  of right singular vectors contains  $K$  deformation bases. Each set of  $D$  rows in the motion matrix  $\mathbf{M}$  of left singular vectors is rearranged as if it were an outer product of rotation coefficients and deformation, and factored as such via a second round of rank-1 SVDs. The rotations and shape/deformation matrix are then affine corrected

25

as described below for rigid objects. That method assumes that the first SVD leaves the singular vectors consistently signed and ordered by deformation and dimension, which is generally not the case.

- 5 If the data to be factorized are derived from a video of the talking head 103, then the first four channels of greatest variation are head height, width, depth, and vertical jaw motion. Therefore, the shape and perhaps the first deformation will be plausible. However, the subsequent ordering of the channels is unpredictable, leading to mutual contamination of the deformation and rotation estimates. A simple example shows how the BHB factorization is vulnerable to less fortuitous data sets and SVDs.

Imagine two beads riding independently on horizontal rails. The bead system has 3D shape and two independent modes of deformation, requiring a shape/deformation basis having 3D modes. Prior art BHB factorization requires rank-9 data to determine shape and two modes of deformation, but the system can only produce tracking data that is rank-5. This means that regardless of the amount of input data, BHB factorization can only recover two modes, i.e., shape and a single deformation, that combine the motion of both beads in way that is not necessarily physically valid. The misordered singular vectors also lead to incorrect rotation estimates, which contaminate subsequent deformation bases with torsions. Therefore, improvements in 3D non-rigid model are needed, as described now in detail.

## Determining Image Correspondences

As shown in Figure 2, step 200 determines the image correspondences 111, i.e., the “evidence” measured in the video. As input, the step takes the video 101 and the locations of the image features 105 identified in the reference image as described above. The image intensity gradients in small regions, e.g., 5x5 pixels centered on the feature locations are sampled 210 spatially and temporally to obtain intensity matrices  $\mathbf{X}$  211 and  $\mathbf{Y}$  212. These matrices 211-212 can be used to determine an optical flow in the video as described below with reference to Figure 5.

The intensity variation matrix  $\mathbf{Y}$  is rank reduced 220 and divided by  $\mathbf{X}$  to obtain the estimated locations 221 of the features in all images and the temporal intensity gradients 222 thus made consistent with 3D non-rigid object motion in the video 101. The rank reduction, as described in detail below is performed with two successive singular value decompositions (SVDs). The locations and associated intensity gradients form the correspondences 111 used by the factorization step 300.

## Factoring Non-Rigid Correspondences

The factorization 300 of the correspondences 111 is described with reference to Figure 3. Factoring uncertain tracking data for rigid object was described by Irani et al. “*Factorization with uncertainty*,” Proc. ECCV, 2000. My factorization corrects some minor errors and then generalizes to non-rigid motion and varied uncertainty structures in the video 101. I use Irani’s variable names to facilitate comparison with prior art.

A *D-interleave matrix*  $\mathbf{E}$  is defined. Post-multiplication with  $\mathbf{E}$  rearranges a matrix with columns representing interleaved data, e.g.,  $xyzxyzxyz$ , to a grouped form, e.g.,  $xxxxxyyyzzz$ ; post-multiplication with  $\mathbf{E}^T$  does the reverse. I use the matrix  $\mathbf{E}$  to rearrange the block-diagonal inverse covariance spatial matrix  $\mathbf{X}$  to form a striped matrix  $\mathbf{X}' = \mathbf{E}^T \mathbf{X} \mathbf{E}$ .

I eigen-decompose  $\mathbf{X}'$  and determine from this a certainty warp  $\mathbf{Q}'$  such that  $\mathbf{Q}' \mathbf{Q}'^T = \mathbf{X}'$  and  $\mathbf{Q}'^T \mathbf{Q}' = \text{eigenvalues}(\mathbf{X}') \mathbf{Q}'$  that maps the directionally weighted least-squares problem implied by  $\mathbf{X}'$  onto an equivalent ordinary least-squares problem where uncertainty in the measurements is now identical and independently distributed (i.i.d).

I split the correspondences into new matrices  $\mathbf{U}_{F \times N}$ ,  $\mathbf{V}_{F \times N}$ , and, perhaps, an optional  $\mathbf{W}_{F \times N}$  containing horizontal, vertical, and optional depth ordinates for  $N$  features in  $F$  frames. I desire a factorization into the shape matrix  $\mathbf{S}_{3K \times N}$  and motion matrix  $\mathbf{M}_{D \times 3K}$  satisfying  $\mathbf{M} \mathbf{S} = [\mathbf{U}, \mathbf{V}, \mathbf{W}]^{(F, N)} = [\mathbf{U}^T, \mathbf{V}^T, \mathbf{W}^T]^T$ , with any residual having minimum Mahalanobis (elliptical) error with respect to the metric defined by  $\mathbf{X}$ . The notation  $\mathbf{M}^{(a, b)}$  signifies a block-transpose where the matrix  $\mathbf{M}$  is transposed by treating each block of  $a \times b$  elements as one element.

I rearrange the correspondences in the new matrices into a horizontally stacked matrix  $[\mathbf{U}, \mathbf{V}, \mathbf{W}]$  in which each row describes one image of the video. This places all variables whose uncertainty is correlated in the same row so that a certainty warp reduction in rank can be applied. The *target* factorization can be rewritten as

$$[\mathbf{U}, \mathbf{V}, \mathbf{W}] \mathbf{Q}' = \mathbf{M} (\mathbf{I}_D \otimes \mathbf{S}) \mathbf{Q}'.$$

I perform the 3D rank- $K$  singular value decomposition 320

$$\mathbf{H}_1 \Delta_1 \mathbf{G}_1 \leftarrow [\mathbf{U}, \mathbf{V}, \mathbf{W}] \mathbf{Q}'$$

to suppress noise under a Mahalanobis error norm specified by the matrix  $\mathbf{X}'$ .

5

I unwarp 330 the left and right factor of the SVD 320 to remove bias introduced by the certainty warp 310 with  $\mathbf{Q}'$ , using a smaller SVD:

$$\mathbf{H}_2 \Delta_2 \mathbf{G}_2 \leftarrow \Delta_1 \mathbf{G}_1^T / \mathbf{Q}'$$

to obtain  $\mathbf{H} \leftarrow \mathbf{H}_1 \mathbf{H}_2 \Delta_2^{1/2}$  and  $\mathbf{G} \leftarrow \mathbf{G}_2 \Delta_2^{1/2}$

Now,  $\mathbf{H}\mathbf{G} \approx [\mathbf{U}, \mathbf{V}, \mathbf{W}]$  is a minimum squared-error rank-3DK approximation with respect to the error norm of  $\mathbf{X}'$ .

I make  $\mathbf{H}$  and  $\mathbf{G}$  331 consistent with the target factorization by finding an invertible transform  $\mathbf{D}_{3DK \times 3DK}$  such that  $\mathbf{M}^{(F, 3K)} = \mathbf{H}\mathbf{D}^{-1}$ , and  $\mathbf{D}\mathbf{G} = (\mathbf{I}_b \otimes \mathbf{S})\mathbf{Q}'$ . Using the above identity, I note that

$$[\mathbf{U}, \mathbf{V}, \mathbf{W}]^{(F, 3K)} = \mathbf{M}\mathbf{S} = (\mathbf{H}\mathbf{D}^{-1})^{(F, 3K)} \mathbf{S} = (\mathbf{I}_b \otimes \mathbf{H}) (\mathbf{D}^{-1})^{(3Dk, 3K)} \mathbf{S},$$

which implies that the desired transform  $\mathbf{D}$  and shape basis  $\mathbf{S}$  can be recovered directly via the rank-3K decomposition 340

$$(\mathbf{D}^{-1})^{(3Dk, 3K)} \mathbf{S} \leftarrow (\mathbf{I}_b \otimes \mathbf{H}) \setminus [\mathbf{U}, \mathbf{V}, \mathbf{W}]^{(F, N)},$$

effectively giving the matrices  $\mathbf{M}$  and  $\mathbf{S}$  121-122.

In contrast to Irani et al., I correctly unwarp the results of the first SVD to yield a fully constrained solution for  $\mathbf{D}^{-1}$ . In the trivial case of isotropic uncertainty, the transform  $\mathbf{D}^{-1}$  is rank-deficient and contains the best estimate of shape. Depending on the sparsity structure of  $\mathbf{X}'$ , this can be broken into small independent problems, e.g., one per feature.

## Factorization from Intensity Gradients

I can also factor 300 directly from intensity variations, which relates to shape and rotation changes through a matrix  $\mathbf{Y}^T = \mathbf{X}(\mathbf{MS})^{(D)}$ . Equivalently, to use the notation of the uncertain factorization method,  $\mathbf{Y}'\mathbf{E} = [\mathbf{U}, \mathbf{V}, \mathbf{W}]\mathbf{X}'$ . Because the uncertainties in  $\mathbf{Y}'\mathbf{E}$  and  $[\mathbf{U}, \mathbf{V}, \mathbf{W}]$  have covariances  $\mathbf{X}'$  and  $\mathbf{X}'^{-1}$  respectively, their certainty-warped forms are equivalent and interchangeable. This means that the factoring 300 can be applied directly to  $\mathbf{Y}'$  by replacing the left hand side of the target factorization with  $\mathbf{Y}'^T \mathbf{E} \mathbf{Q}' \mathbf{\Lambda}^{-1}$ , where  $\mathbf{\Lambda}$  is the diagonal matrix of eigenvalues of  $\mathbf{X}'$ .

## Central Problem

The crux of the problem is to find an optimal correction matrix  $\mathbf{J}^{-1}_{3K \times 3K}$  that transforms the result  $\mathbf{M}, \mathbf{S}$  of the above factorization into a properly structured motion matrix  $\mathbf{M}$  and the shape basis  $\mathbf{S}$ . Recall from equation 3 that each  $D \times 3$  block in the matrix  $\mathbf{M}$  121 is a scaled rotation whose rows effect the  $x$  and  $y$  image projection of one deformation basis. These rows have an equal norm and are orthogonal. Moreover, they are orthogonal to the  $y$  and  $x$  projections taken from any blocks to the left and right, because these blocks are all generated from the same rotation.

For a rigid object, the exact set of necessary and sufficient norm/orthogonality constraints that the motion matrix  $\mathbf{M}$  must satisfy can be approached by as a squared-error problem via nested least-squares procedures. This is the strategy of the rigid affine correction described above. However, in the non-rigid case, this strategy does not apply because the least-squares procedure first performs a

division that discards information about the quadratic terms that will be needed for a second eigen decomposition.

A direct solution is a very difficult problem. Prior art solutions have centered on finding numerically well-behaved heuristics. For example, the BHB factorization sets  $\mathbf{J} = \mathbf{I}_K \otimes \mathbf{G}$ , a block-diagonal correction that assumes that the SVD correctly arranges all of the information about a deformation basis in the appropriate column-triple in the matrix  $\mathbf{M}$ .

Experiments with projections of measured 3D image data, whose principal components are known, indicate that the correction matrix  $\mathbf{J}$  is dense, particularly above the diagonal. This means that the SVD mixes variation due to minor deformations into the shape and principal deformations. Indeed, it is quite difficult to construct a data set for which the correction matrix  $\mathbf{J}$  has anything vaguely close to a block-diagonal structure, even with vast amounts of noiseless data. Synthetic-data experiments suggest that the scale of the deformations must drop off quadratically in order for the initial SVD to properly group their  $x$ ,  $y$ , and  $z$  components. Even then, it is unlikely that the components are consistently ordered within all groups.

Appendix A describes a family of solutions that generalizing the prior art affine correction procedure to non-rigid objects. However, these solutions still suffer from rank-deficiency problems because the number of unknowns grows quadratically while the rank of the constraints grows linearly because the matrix  $\mathbf{J}^{-1}$  has  $9K^2$  unique elements, while there are  $4K$  non-redundant constraints in the motion matrix  $\mathbf{M}_f$  of each image in the video: one norm and one orthogonality constraint per block, and two orthogonality constraints from the first block to each



block to its right. This is a property of the image-formation model of equation 1. Moreover, these constraints tend to be redundant between frames. Therefore, very large amounts of video are needed to correctly estimate  $\mathbf{J}$  using information in the matrix  $\mathbf{M}$  only. Consequently any procedure based purely on the expected structure of the motion matrix will fail as the number of deformation modes grows.

### Flexible Factorization

My solution constrains the shape/deformation basis  $\mathbf{S}$  so the deformations are as small as possible relative to the mean shape. Then, the displacement of projected features away from the object's centroid are maximally consistent with the object's shape and only residually accounted for by its deformations. Equivalently, whenever possible, the motions of the feature points should be expressed parsimoniously by rigid transforms, i.e., rotations and scale changes, rather than unparsimoniously by combinations of deformations. Otherwise, all motion could be explained as deformations.

Therefore, I multiply the shape/deformation basis  $\mathbf{S}$  by the correction matrix  $\mathbf{J}$  to obtain a corrected shape/deformation matrix, and define a matrix  $\mathbf{Z}$  as  $\mathbf{I}_3 \cdot \text{diag}[0, \mathbf{1}_{1 \times K-1}]$ . Then, I minimize the Frobenius norm of  $\mathbf{ZS}$ , which is the part of the shape/deformation matrix that contains deformations.

I now have two constraints: the structure of the motion matrix  $\mathbf{M}$ ; and the parsimony of the deformations  $\mathbf{S}$ . The problem is that the motion matrix gives constraints on the correction matrix  $\mathbf{J}^{-1}$  via  $\mathbf{M}^* = \mathbf{MJ}^{-1}$ , while the shape/deformation matrix gives constraints on the correction matrix  $\mathbf{J}$  via  $\mathbf{S}^* = \mathbf{SJ}$ .

To work around this inconvenience, I rewrite the motion constraint as  $M\mathbf{J} = \mathbf{M}$ , where  $M$  is an initial estimate of the corrected motion matrix  $\mathbf{M}$ .

To make the initial estimate  $M$ , I can use the non-rigid affine correction described in Appendix A or the BHB factorization, and construct a properly structured motion matrix from the result. Both methods have weaknesses.

## Model Extraction

Therefore, as shown in Figure 4 for a more robust procedure, I first invert the sign of the left singular vectors in the matrix  $\mathbf{M}$  as necessary to minimize a residual of the norm/orthogonality constraint, i.e., to maximize orthogonality:

$$\sum_{\mathbf{M}_f \in \mathbf{M}} \left\| (\mathbf{M}_f^T)^T (\mathbf{M}_f^T)^T - ((\mathbf{M}_f)^T (\mathbf{M}_f) \otimes \mathbf{I}_D / D \right\|_F^2 \quad (4)$$

where  $F$  indicates the Frobenius norm. Maximizing orthogonality leads to better estimates of rotation, and therefore to better estimates of shape. Tests for the desirability of various sign-flips can be done efficiently by caching intermediate results. Short-distance column-swaps can be evaluated in the same manner. The same operation is done for the shape basis  $\mathbf{S}$  to yield the inverted matrices  $\mathbf{M}'$ - $\mathbf{S}'$ .

I then affine-correct each column triple of the inverted matrices as described for affine correction, and 3D-rotate each column-triple to a common coordinate frame. I then stack all the column-triples, and determine an affine correction  $\mathbf{G}^{-1}$  as described above, and apply the affine correction to the stacked column-triples of the matrices.

Here, unlike the BHB procedure, each column-triple has a unique correction. Also, by eliminating the information-lossy factorization into  $\mathbf{R}_f$  and  $\mathbf{c}_f$ , I obtain estimates of  $\mathbf{G}^{-1}$  that better reduce the error in the model as set out in equation 1. For each transform to the motion matrix  $\mathbf{M}$ , a compensatory inverse transform is applied to the shape basis  $\mathbf{S}$ .

For each image (frame), I then factor each  $\mathbf{M}_i \in \mathbf{M}$  into the rotation matrix  $\mathbf{R}_i$ , and the deformation coefficients  $\mathbf{c}_i$  using orthonormal decomposition constraints using a procedure based on Horn's absolute orientation method, see Horn, "Closed form solution of absolute orientation using unit quaternions," J. Opt. Soc. Am. A, Vol. 4, pp. 629-642, 1987. That method directly factors a matrix into a rotation and a vector. I then reconstitute the properly constructed motion matrix  $\mathbf{M}$  using equation 3. Like the BHB procedure, this assumes a block-diagonal correction. However, my estimate is only an initial estimate of the motion matrix, and the off-diagonal elements of the correction matrix  $\mathbf{J}$  are recovered as described below. It is also possible to factor each into a rotation and deformation coefficients via rank-1 SVD, then affine-correct the rotations yet again.

Combining the constraints from the motion and shape/deformation matrices, I obtain the objective function

$$\min_{\mathbf{J}} \text{trace}((\mathbf{M}\mathbf{J} - \mathbf{M})^T - (\mathbf{M}\mathbf{J} - \mathbf{M})) + \text{tr}(\mathbf{S}^T \mathbf{J}^T \mathbf{Z} \mathbf{J} \mathbf{S}).$$

(5)

This objective function seeks the correction operator  $\mathbf{J}$  that yields the expected structure in the motion matrix  $\mathbf{M}$  with the *smallest* possible deformations in the

shape/deformation basis  $\mathbf{S}$  and coefficients matrix  $\mathbf{C}$ . The error is minimized by a solution to a system of equations  $\mathbf{MJ} = \mathbf{M}$  and  $\mathbf{ZJS} = \mathbf{0}$ . The sought correction matrix  $\mathbf{J}$  is obtained from a sparse right-handed division

$$\mathbf{J} \leftarrow \frac{\begin{bmatrix} \mathbf{I}_{3K} \otimes \mathbf{M} \\ \mathbf{S}^T \otimes \mathbf{Z} \end{bmatrix}}{\begin{bmatrix} \mathbf{M} \\ \mathbf{0}_{3KN \times 1} \end{bmatrix}}. \quad (11)$$

This allows me to determine  $\mathbf{S} \leftarrow \mathbf{JS}$ , and  $\mathbf{M} \leftarrow \mathbf{M/J}$ . Because equation 11 uses information on both sides of the SVD, it is well constrained. In practice, I find that the upper triangle and several sub-diagonals of  $\mathbf{J}$  are usually dense, indicating that information about any particular deformation is indeed spread over several columns of the motion matrix  $\mathbf{M}$ .

It is also possible to repeatedly apply equation (11) after re-estimating rotations and deformation, but such iterations only yield small improvements in the shape/deformation bases  $\mathbf{S}$ .

The estimate rotation matrix and shape matrix can be refined by projection onto a manifold of orthogonal matrices. The refined matrices are then solved to obtain the deformation coefficients  $\mathbf{c}$ , see U.S. Patent Application Sn., 09/791,117 “*Modeling shape, motion, and flexion of non-rigid 3D objects in a sequence of images*,” particularly equations 6 and 8 for details.

## Image Gradients

As stated earlier, the method 100 according to the invention can be entirely performed with image intensity gradients when considering the setting as shown in Figure 5. A small region  $R$  510, e.g.,  $5 \times 5$  pixels, in an image  $I_0$  511 shifts to a new location in a next image  $I_1$  512. Assuming that the images view a constantly illuminated Lambertian (diffuse reflective) surface, the 2D optical flow  $\mathbf{f}_{2 \times 1}$  500 can be estimated, to the first-order, from the spatial image gradient

$$\nabla_p = dI_0(\mathbf{p})/d\mathbf{p} \text{ as } \mathbf{f} \leftarrow \mathbf{X}\mathbf{y}, \quad (6)$$

where the  $\mathbf{X}_{2 \times 2} = \int_R \nabla_p \nabla_p^T d\mathbf{p}$  is the local  $R$  510 spatial variation of the image intensity within the image  $I_0$ , and  $\mathbf{y}_{2 \times 1} = \int_R I_0(\mathbf{p}) - I_1(\mathbf{p})$  is the temporal image intensity variation between images  $I_0$  and  $I_1$ . For a given video, good estimates of  $\mathbf{X}$  are usually possible. However, the temporal variation  $\mathbf{Y}$  is sensitive to noise in the image intensities. Assuming this noise is Gaussian distributed, the spatial variation  $\mathbf{X}$  has special significance as the inverse covariance matrix of the flow estimate  $\mathbf{f}$ . The eigen-vectors of  $\mathbf{f}$  give the directions in which the optical flow is most and least certain.

I represent  $N$  local flows for each of the  $F$  images simultaneously in matrices  $\mathbf{F}_{DN \times F}$   $\mathbf{Y}_{DN \times F}$ , and diagonally stacked  $\mathbf{X}_{DN \times DN}$ . If  $\mathbf{X}$  describes the spatial intensity variation around the features in the reference image 102, and each column of each matrix  $\mathbf{Y}_f \in \mathbf{Y}$  describes temporal intensity variation between the reference image  $I_0$  and a target image  $I_f$ , then in the absense of additional constraints,  $\mathbf{Y} = \mathbf{X}\mathbf{F}$ . The covariance of the uncertainty in  $\mathbf{F}$  is  $\Sigma_F = \mathbf{X}^{-1}$ , and conversely  $\Sigma_Y = \mathbf{X}\Sigma_F\mathbf{X}^T = \mathbf{X}$ .

I will now described how all of the operations of the above section can be applied to  $\mathbf{X}$  and  $\mathbf{Y}_f$ . First, I eigen-decompose  $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \leftarrow \Sigma_{\mathbf{Y}}$ , and use  $\mathbf{Q} = \mathbf{\Lambda}^{1/2}\mathbf{V}^T$  for certainty-warped operations on  $\mathbf{Y}$ . A certainty warp affine-transforms a problem having an elliptical (Mahalanobis) error norm so that minimum Mahalanobis-error solutions are obtained from least-squares procedures such as matrix division and SVD. I use this to estimate pure translations

$$\mathbf{T} \leftarrow (\mathbf{Q}\mathbf{X}(\mathbf{1}_{n \times 1}\mathbf{I}_D))^+(\mathbf{Q}\mathbf{Y}). \quad (7)$$

This is a certainty-warped calculation of the mean displacements. The pseudo-inverse denoted by the superscripted plus sign is determined by using QR-decomposition and inversion of the upper-triangular  $D \times D$  matrix. I now remove translation and incorporate position into the temporal intensity variations, obtaining

$$\mathbf{Y}' = \mathbf{Y} + \mathbf{X}(\mathbf{P}_0 - \bar{\mathbf{P}}_0 - \mathbf{T}), \quad (8)$$

where  $\bar{\mathbf{P}}_0$  is the centroid. Here, I use the convention that when adding or subtracting two matrices of different size, the smaller matrix is tiled to match the size of the larger matrix. The matrix  $\mathbf{Y}'$  is now a function of rotations and deformations only, satisfying

$$\mathbf{P} = (\mathbf{X}\mathbf{Y}')^{(D)} = \mathbf{M}\mathbf{S} \quad (9)$$

How to factor the zero-meaned correspondence estimates  $\mathbf{P}$  with respect to their uncertainty, i.e., covariance  $\Sigma_{\mathbf{xy}} = \Sigma_{\mathbf{F}} = \mathbf{X}^{-1}$ , into  $\mathbf{M}$  and  $\mathbf{S}$ , is described above, and how to do the same factoring directly from intensity variations  $\mathbf{Y}'$  with respect to their uncertainty  $\Sigma_{\mathbf{Y}'} = \mathbf{X}\Sigma_{\mathbf{F}}\mathbf{X}^T = \mathbf{X}$ . Then, the flexible factorization 300 described above applies directly.

### 3D Optical Flow

I use the fact that non-rigid motion is a low-rank multi-linear process to concurrently *track* a 3D *non-rigid* object and *recover* its 3D shape and deformation basis simply by *manipulating the rank of optical flow computations*.

Irani, in “*Multi-frame optical flow estimation using subspace constraints*,” Proc. ICCV, 1999, describes a low-rank multi-linear process that allows one to track a surface of a 3D rigid object. Her process was based on a premise that the optical flow and associated temporal image gradients from a reference image to subsequent images are bilinear products of two matrices whose low rank is deduced from the camera and scene type.

However, until now, no solutions have been available for the more complex problem of concurrently tracking, and acquiring shape and deformation of non-rigid objects. The fact that the present method provides a low-rank multi-linear process for non-rigid objects enables the rendering of deformable 3D models recovered from deformable objects viewed in videos.

The method according to the invention make it possible to concurrently track a 3D non-rigid object **and** acquire its 3D shape and deformations by manipulating the **rank** of the optical flow calculations. The present method reduces  $\mathbf{P}$  to a rank- $3K$ . This force the motion data to be consistent with the subspace of plausible non-rigid 3D models. Moreover, because temporal intensity gradients  $\mathbf{Y}$  are locally linearly in motion  $\mathbf{P}$ , i.e.,  $\mathbf{Y} = \mathbf{X}\mathbf{F} = \mathbf{X}(\mathbf{P}_0 - \bar{\mathbf{P}}_0 - \mathbf{T})$ , uncertainty-informed rank-reduction of the temporal intensity variation matrix  $\mathbf{Y}$  similarly constrain the optical flow to be in the same subspace.

The key to my method is to manipulate  $\mathbf{Y}'$  in equation 8 so that the rank constraints implied by equation 9 are applicable. This is accomplished by the intensity-based flexible factorization 300 as described above. I also provide a more efficient  
5 procedure.

As described above with reference to Figure 2, I begin by choosing the reference image  $I_0$  102 from approximately the middle of the video 101. Then, I determine the spatial variation  $\mathbf{X}$  221 of the optical flow from image regions 501 within  $I_0$ , and the temporal variation  $\mathbf{Y}'$  221 of the optical flow from comparisons of those regions to similarly locate corresponding regions in all other images of the video 101.

Since  $\mathbf{MS}$  has rank  $3K$ , equation 9 tells me that  $\mathbf{Y}'$  has rank  $3DK$ . I eigen-decompose  $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \leftarrow \Sigma_{\mathbf{Y}'} = \mathbf{X}$ , and use  $\mathbf{Q} = \mathbf{\Lambda}^{1/2}\mathbf{V}^T$  to certainty-warp the uncertainty-informed thin SVD  $\mathbf{U}\mathbf{\Gamma}\mathbf{W}^T \leftarrow \Sigma_{\mathbf{Y}}$ . This is the SVD 320 in Figure 2.

Because  $\mathbf{Q}^T\mathbf{Q} = \mathbf{X}^{-1}$ , the product  $\mathbf{Q}^T\mathbf{U}\mathbf{\Gamma}\mathbf{W}^T \approx \mathbf{X}^{-1}\mathbf{Y}' \approx \mathbf{X} \backslash \mathbf{Y}$  is an uncertainty-informed rank- $3DK$  approximation of the rotational and deformational component  
20 of the inter-image correspondences. This is the division in step 220.

To conform with  $\mathbf{MS}$ , rearranging the product yields the final rank-reduction to rank  $3K$ ,  $\mathbf{U}'\mathbf{\Gamma}'\mathbf{W}'^T \leftarrow (\mathbf{Q}^T\mathbf{U}\mathbf{\Gamma}\mathbf{W}^T)^{(D)}$ . Finally, I restore the translations to obtain correspondences  $\mathbf{P}_{NxDF} = \mathbf{U}'\mathbf{\Gamma}'\mathbf{W}'^T + \mathbf{T} + \bar{\mathbf{P}}_0$ . New temporal image gradients are  
25 sampled 210 with respect to these correspondences, and the process repeats until it converges.



Unlike Irani's process, my method handles rank-reduced optical flow for non-rigid scenes and objects, properly certainty-warps the intensity variations with respect to their own uncertainty *prior* to SVD, and rank constraints that are *exact* because they are inherited directly from the forward model; and the SVD produces results that are useful beyond rank-reduction as follows.

The SVD yields  $\mathbf{M} \leftarrow \mathbf{U}' \sqrt{\mathbf{\Gamma}'}$  and  $\mathbf{S} \leftarrow \sqrt{\mathbf{\Gamma}'} \mathbf{W}'$ . These results can be used to extend the range of the video frames being tracked by predicting correspondences in newly generated images via linear extrapolation of the rows at the top and bottom of the motion matrix  $\mathbf{M}$ . The extrapolations can also be used to warp the video to create new frames that continue the observed motion.

The factorization constraints the search for correspondences, and the search provides information for the factorization. As the process analyzes more and more video frames, the space of possible non-rigid 3D models becomes increasingly more robust. When the correspondence process converges, I need only determine the correction matrix  $\mathbf{J}$  to recover the final shape/deformation basis  $\mathbf{S}$  and the motion matrix  $\mathbf{M}$ . Various efficiency improvements can be obtained through incremental SVD (so that prior SVD results are re-used), multi-scale search, and using a subset of all images. For very long sequences, one can analyze a subsequence of the video to obtain the shape/deformation matrix  $\mathbf{S}$ . This matrix can then be used for robust, efficient tracking in the rest of the video.

Figure 6 show motion parameters recovered from an unconstrained video, i.e., there are no "markers" on the object from which the model is recovered. The

object is a front-facing “talking head” with a relatively low texture, lighting changes during the sequence, the camera parameters are unknown, and there is motion in the background. The recovered parameters include translation 601, rotation 602, scale 603, and deformation coefficients 604. It is worth noting that the prior art Irani rank-reduced flow method fails to track this sequence, even when given the same rank constraints as here. The video was also presented to the prior art BHB process, which failed to separate jaw motions from head rotations.

Figure 7 compares the RMS error of the present flex factoring 701 with the prior art (BHB factoring) 702. The improved reduction in the residual is obvious for the factoring according to the invention.

### **Effect of the Present Recovery Method**

I have described a linear method for recovering 3D shape, motion, and articulations of non-rigid 3D objects from low-quality low-resolution video such as low-end consumer videos having partially degenerate textures. The output of my method is a 3D deformable model and 3D motion/morph parameters. I have identified a “parsimony” constraint, and use this constraint to correct singular valued decompositions of videos with undesirable properties. The recovered model and parameters can be used to synthesize novel videos. The invention is particularly useful for animating, altering, or morphing characters in movies and computer games. Current practice tediously extract models by using markers and manual extraction methods, on a frame-by-frame basis. This is frequently done in the movie industry when stand-ins or stunt-workers take place of the real actors, and it is later desired to overlay the stand-ins with images of the actors. My

methods allow one to automate these processes at great savings in time, effort, and cost.

5 This invention is described using specific terms and examples. It is to be understood that various other adaptations and modifications may be made within the spirit and scope of the invention. Therefore, it is the object of the appended claims to cover all such variations and modifications as come within the true spirit and scope of the invention.

099306 080841  
F08080 9E052560